CHALMERS
UNIVERSITY OF TECHNOLOGY

From Ad-Hoc Data Analytics to DataOps

From Ad-Hoc Data Analytics to DataOps

Aiswarya Raj Munappy
Chalmers University of Technology
Göteborg, Sweden
aiswarya@chalmers.se

David Issa Mattos
Chalmers University of Technology
Göteborg, Sweden
davidis@chalmers.se

Jan Bosch
Chalmers University of Technology
Göteborg, Sweden
jan.bosch@chalmers.se

Helena Holmström Olsson
Malmö University
Malmö, Sweden
helena.holmstrom.olsson@mau.se

Anas Dakkak
Ericsson
Stockholm, Sweden
anas.dakkak@ericsson.com

ABSTRACT

The collection of high-quality data provides a key competitive advantage to companies in their decision-making process. It helps to understand customer behavior and enables the usage and deployment of new technologies based on machine learning. However, the process from collecting the data, to clean and process it to be used by data scientists and applications is often manual, non-optimized and error-prone. This increases the time that the data takes to deliver value for the business. To reduce this time companies are looking into automation and validation of the data processes. Data processes are the operational side of data analytic workflow.

DataOps, a recently coined term by data scientists, data analysts and data engineers refer to a general process aimed to shorten the end-to-end data analytic life-cycle time by introducing automation in the data collection, validation, and verification process. Despite its increasing popularity among practitioners, research on this topic has been limited and does not provide a clear definition for the term or how a data analytic process evolves from ad-hoc data collection to fully automated data analytics as envisioned by DataOps.

This research provides three main contributions. First, utilizing multi-vocal literature we provide a definition and a scope for the general process referred to as DataOps. Second, based on a case study with a large mobile telecommunication organization, we analyze how multiple data analytic teams evolve their infrastructure and processes towards DataOps. Also, we provide a stairway showing the different stages of the evolution process. With this evolution model, companies can identify the stage which they belong to and also, can try to move to the next stage by overcoming the challenges they encounter in the current stage.

CCS CONCEPTS

• **Software and its engineering**;

KEYWORDS

DataOps, Data Pipelines, Continuous Monitoring, DevOps, Data technologies, Agile Methodology

ACM Reference Format:
Aiswarya Raj Munappy, David Issa Mattos, Jan Bosch, Helena Holmström Olsson, and Anas Dakkak. 2020. From Ad-Hoc Data Analytics to DataOps. In *International Conference on Software and Systems Process (ICSSP '20), October 10–11, 2020, Seoul, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3379177.3388909

ICSSP '20, October 10–11, 2020, Seoul, Republic of Korea

ACM ISBN 978-1-4503-7512-2/20/10…$15.00
https://doi.org/10.1145/3379177.3388909

1 INTRODUCTION

Data is the key asset for organizations as it helps in better decision making, analyze performance and solving problems, to analyze the consumer behavior and market and so on. Moreover, data is the backbone for many hot and trending technologies like machine learning and deep learning [16]. The increased importance of data leads to the acquisition and storage of data in higher volumes which in turn gave rise to fields like Big Data, data mining and data warehousing. Data being the fuel for the digital economy, the need for data products like machine learning datasets, dashboards and visualizations is tremendously increasing. Organizations invest in data science and data analytics to solve problems with the collected data. Organizations realize that data is the key factor of success and as a result, they invest an enormous amount of money in the development of data products [17]. Data products are built through a sequence of steps called data life cycle wherein for each step there will be both hardware and software requirements. Consequently, it is very essential to find the right balance of investment in requirements in different stages of the data life cycle [17]. Data management, data life cycle management, data pipeline robustness, fast delivery of high-quality insights are some of the major data problems that prevent companies from achieving their full potential.

DevOps is a set of practices that helps to build a collaboration between software development and information technology operations which in turn reduces the software development lifecycle and helps in continuous and fast delivery of high-quality systems. Thus, it is a methodology adopted in Software Engineering to aid agile software development [25]. Agile methodology focuses on empowering individuals, rapid production of working software, close collaboration with customers and quick response to the change in customer requirements [30]. Agile development is directly facilitated by CI/CD practices because it aids in software changes

reaching production more frequently and rapidly. Consequently, customers get more opportunities to experience and provide feedback on changes [23].

Industries apply agile methodology, DevOps and CI/CD methodologies in software development. Data being an artifact like code, data analytics can also be benefited by the application of best practices of these methodologies in data analytics. DataOps is a process-oriented methodology that is derived from DevOps, continuous integration/continuous delivery and agile methodology for the quick delivery of high-quality insights to the customers. Introduction of agile development, CI/CD methodologies, and DevOps paves way for collaborative working, faster fixes, increased team flexibility, agility, cross-skilled and self-improving teams.

Many companies have succeeded in implementing DevOps, agile and CI/CD practices in their organization. However, there are only a few companies that have succeeded in adopting DataOps practices. In order to advance the concept of agile development and CI/CD and move towards DataOps, there are several steps that need to be taken. These several steps taken by the companies form a stairway and contributes to the evolution model of DataOps. Although it resembles DevOps practices, applying the same practices in Data Analytics is quite challenging as both of these disciplines are unique in their own respect and the skill-set, interest of practitioners involved in Data Analytics are very different from the people who are involved in Software development. Therefore, the challenges faced by companies at each stage of progression towards DataOps will be much different from challenges associated with the evolution of DevOps.

The contribution of this paper is three-fold. First, it analyses the various definitions of DataOps from the literature as well as from the interviewers and then derives a definition for DataOps including the main components identified. Second, based on a case study with a large mobile telecommunication organization, we analyze how multiple data analytic teams evolve their infrastructure and processes towards DataOps. Third, we create a stairway of the evolution process. DataOps is a recently coined term, it is important to understand how companies are progressing towards DataOps. The evolution model demonstrates the essential requirements to climb a step in the stairway and also lists the set of challenges encountered while moving from one stage to the next.

The rest of this paper is organized into six sections. Section II is a description of the background and related work. In section III, the research methodology adopted for conducting the study is introduced. Section IV focuses on the findings of the case study, framing the definition for DataOps and the evolution stages. Section V details threats to validity and finally section VI summarises our conclusions and completes this paper.

2 RELATED WORKS

The peer reviewed works related to DataOps are quite few in number. Ereth [20] in his paper discussed a working definition for DataOps. Sahoo et. al presented a study which compares DataOps to DevOps and outlined the DataOps process and platform as well as the data challenges in manufacturing and utilities industries.

According to Julian Ereth [20], DataOps is a collection of various practices and technologies, than a particular method or tool. His study has resemblance with the first part of this paper where a definition for DataOps is derived from the literature as well as from the practitioners' understanding. Using a multi-vocal literature review (MLR) approach supplemented by interviews, the author analyzed and derived a definition for an ambiguous concept "DataOps". The author has also developed a framework that differentiates between the exploration of DataOps as a discipline, which includes methods, technologies and concrete implementations, and the investigation of the business value of DataOps. However, the paper does not discuss how DataOps is different from Big Data Analytics, DevOps or CI/CD approach.

P. R. Sahoo et. al defines DataOps as an application of DevOps to data and they draw a parallel between DataOps and DevOps concepts. The authors define DataOps as DevOps for data analytics which eliminates inefficiencies, creates opportunities for collaboration, and promotes reusability to reduce operational costs. The study highlights how DataOps can be used in the data analytics discipline to bring revolutionary changes to business [29]. Also, it identifies the six significant steps of the DataOps process such as business requirements planning, data acquisition, data transformation, data repository management, data modeling, and insight publication.

Previous studies on data-driven development [19] describe the way companies evolve through their ability to use data. The study shows that companies follow a predictable pattern and start with an ad-hoc and manual approach to a data-driven approach. The authors developed a stairway with evolution stages. The first stage is the ad-hoc data collection. Challenges with manual data collection lead to automated data collection, followed by the introduction of dashboards that automatically updates with data from the field. After this stage, due to the constant flow of new insights evolving dashboards are introduced. Eventually, data-driven decision making is adopted for everything including sales, performance reviews, hiring and other processes. This work has a close resemblance to our study as it deals with data and evolution phases.

3 RESEARCH METHODOLOGY

The goal of this study was to formulate a definition for DataOps and to identify the phases of DataOps evolution.

3.1 Setting the RQs

The RQs defined in the study are as given below:-

- **RQ1.** How do practitioners define "DataOps"?
- **RQ2.** What are the different maturity stages Ericsson has gone through while trying to evolve from ad-hoc data analysis to DataOps?

To set the basic understanding of DataOps concepts and the essential components, we adopted the Multi-Vocal Literature Review approach following the instructions given by [22]. Then we conducted an interpretive single-case study, following the guidelines by [28], to acquire a deeper understanding of the data analytic approach followed at Ericsson. The main focus of this study is to understand and explain how the DataOps approach is perceived by Data Scientists, Data Analysts and Data Engineers to shorten the end to end data analytic life-cycle time and to enable collaboration. The impediments identified at each phase are based on

our interpretations of the experiences of experts who work with data in a real-time scenario with real-world data collected from edge devices. The multiple cases from different teams in the same company are used in this study because it facilitates the exploration of a particular concept in a real-life setting as well as through a variety of lenses [18]. The overall research design and major steps in the process of the study are described below.

3.2 Multi-Vocal Literature Review

An MLR is a form of a Systematic Literature Review (SLR), which includes the Grey literature in addition to the published literature (e.g., journal and conference papers) [21]. Grey literature in SE can be defined as any material about Software Engineering that is not formally peer-reviewed nor formally published. The multi-vocal literature review approach was selected because it allowed us to gain more understanding of DataOps practices. As explained in [22], we analyzed if there is a great potential for benefiting from grey literature in the DataOps study and we identified that clearly, this approach is the best-suited one for studying DataOps. Because, the formal literature on the other hand DataOps is highly limited and on the other hand, there are quite several blogs, video media, and technical reports. Moreover, MLRs are useful since they can provide summaries of both the state-of-the-art and practice in a given area. We searched the academic literature using the Google Scholar, IEEE Xplore, ACM digital library and the grey literature using the regular Google search engine.

3.3 Need for MLR:

To learn more about the concept of DataOps, we did an initial search for the formal academic literature in different databases such as Google Scholar, IEEE Explore, ACM digital library, Web of Science, Scopus and ScienceDirect. However, we could not find a considerable number of peer-reviewed papers on the topic. Consequently, we decided to conduct a Multivocal Literature Review, based on all available literature on a topic.

According to Ogawa et.al a broader view about a particular topic can be obtained by using this wide spectrum of literature as they include the voices and opinions of academics, practitioners, independent researchers, development firms, and others who have experience on the topic [27].

Garousi et al. state that the practitioners produce literature based on their experience, but most of them are not published as academic literature. Also, the voice of the practitioners better reflects the important current state-of-the-art practice in SE. Therefore, it is important to include Grey literature too in the systematic review [21].

3.4 Process of MLR

The Multi-vocal literature review procedure adopted for the study is demonstrated in Fig. 1. The systematic review employs a string-based database search to select relevant studies from the literature. All retrieved literature was exported to MS Excel for further processing. The exported references were screened based on inclusion-exclusion criteria. The inclusion and exclusion criteria considered in our study are as shown below.

- **Inclusion Criteria :**
  (1) Papers and Google links describing the steps of the DataOps approach, essential components of DataOps, benefits, and challenges.
  (2) Papers describing the Big data pipelines, Big data processing pipelines
- **Exclusion Criteria :**
  (1) Duplicates and non-English

Table 1: Description of use cases and roles of the interviewees

Case	Use cases at Ericsson	Interviewed Experts	
		ID	Role
A	Automated data collection for data analytics	R4	Senior Data Scientist
B	Building data pipelines	R1	Integration and Operations Professional
C	Toolkit for Network Analytics	R2	Analytics System Architect
D	Building CI pipelines for Data Scientist team	R7	Data Scientist
E	Tracking the Software Version	R5	Senior Customer Support Engineer
F	Testing the Software Quality	R6	Developer Customer Support
G	KPI Analysis Software	R3	Senior Data Engineer
H	Building data pipelines for CI and CD data	R8	Program Manager

3.5 Exploratory Case study

The study was conducted in collaboration with Ericsson. Ericsson is a Swedish multinational network and telecommunications company. The company provides services, software, and infrastructure in information and communications technology. The objective of the study is to explore the essential stages of the Data Analytic approach which Ericsson follows in their real-world settings and also to investigate its similarity to the popular DataOps approach. Each case in the study refers to a team at Ericsson working with the data they collect from different sources. For the study, a sample pool of Data Scientists, Data Analysts and Data Engineers were selected by one of the authors according to their expertise in the area of Data Analytics. Selected practitioners were invited to participate in the interview study and 4 of them showed interest to participate. After the interviews, interviewees were asked to suggest the names of their colleagues whom they think will be potentially interested in the study. Invitations were sent out to them as well and 4 of them participated in the study thus making a total of 8 interviews. Table 1 illustrates the role of our interviewees and the use cases.

3.6 Data Collection

Empirical data was collected through semi-structured interviews. Based on the objective of the research, to explore the data analytic approach employed at Ericsson, an interview guide with 45 questions categorized into six sections was formulated. The first and second sections concentrated on the background of the interviewee. The third and fourth sections focused on the data collection and processing in various use-cases and the last section inquired in

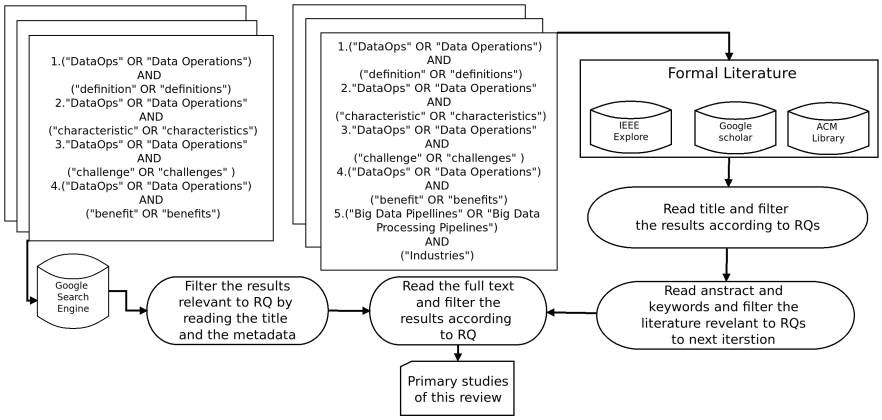

Figure 1: Multi-vocal literature review procedure applied in the study

detail about data testing and monitoring practices and the impediments faced during every phase of the data pipeline. The interview guide was prepared by the first author and was reviewed by all the other authors. Based on the comments and recommendations some additional questions were added, a few similar questions were merged and some irrelevant questions were removed forming an interview protocol with 30 questions spread across six different categories. All interviews were conducted via video conferencing except for three which were done face-to-face and each interview lasted 50 to 100 minutes. All the interviews were recorded with the permission of respondents and were transcribed later for analysis.

One of the authors of this paper is an Ericsson employee who works quite a lot with the data teams. The first two authors of this paper are consultants at Ericsson and attend weekly meetings with Data Scientists and Data Analysts. Data collected through the meetings and discussions are also incorporated. The contact points at Ericsson were also a great help while validating the collected data.

3.7 Data Analysis

After the interviews, audio recordings of the interview were sent for transcription and a summary of each interview was prepared by the first author highlighting the important focus points of the interview. The investigated points from the summary were cross-checked several times with the audio recordings and interview transcripts obtained after transcription. A theoretical thematic data analysis approach was selected for coding [26]. The first author coded each relevant segment of the interview transcript in NVivo. For the first iteration, the objective was to identify the use-cases discussed by each interviewee and phases of data analytics used by their team. After identifying the phases, a second iteration was performed to investigate the impediments encountered to completely set up DataOps practices at Ericsson. Thematic coding was performed by setting high-level themes as (i) Data Collection, (ii) Data Analytics, (iii) DevOps, (iv) Automation, (v) Data Testing, (vi) Data monitoring, (vii) Agile development. After careful analysis of the collected data, the first two authors agreed on the presentation of results in the paper. From the analysis, results were tabulated and sent to the other authors for collecting their reflections and then the final summary of the cases and results were sent to the interviewees for validation.

4 FINDINGS

This section presents a definition of DataOps derived from literature as well as from the definitions given by experts during the interview study. Based on the study, we have constructed a five-stage evolution model of data strategy adopted at Ericsson to meet the evolving requirements of the customer. Our study is carried out with eight use cases as mentioned above. Requirements for moving from one step to the next and impediments encountered at each phase are identified and described as following.

4.1 Definition of DataOps

The exploratory case study and interviews show that different practitioners have different understandings about DataOps. During the interview, practitioners defined DataOps as "a process which fills the gap between data and operations team", "an efficient way of managing the activities in the entire data life cycle", "a method to showcase the interdependence of end to end data analytic process" or "an approach to eliminate data silos by connecting different data pipelines“.

Similarly, there are several definitions for DataOps in the grey literature. The concept of DataOps was first introduced by Lenny Liebmann in his blog post titled “3 reasons why DataOps is essential for big data success.” in 2014 [1]. However, it got popularity in 2015 through the blog post “From DevOps to DataOps” [**?**] by Andy Palmer. Andy Palmer described DataOps as a discipline that “addresses the needs of data professionals on the modern internet and inside the modern enterprise” [12]. Gartner’s glossary defines DataOps as "hub for collecting and distributing data, with a mandate to provide controlled access to systems of record for customer and marketing performance data, while protecting privacy, usage restrictions and data integrity" [2]. There are several other definitions like DataOps "spans the entire analytic process, from data acquisition to insight delivery“ [4], "is a better way to develop and deliver analytics" [8], “is a new way of managing data that promotes communication between, and integration of, formerly siloed data, teams, and systems” [5] or "is illustrated as intersecting Value and

Innovation Pipelines" [7]. From the above definitions, it can be seen that many of the experts define DataOps as an end-to-end process spanning from data acquisition to the insight delivery.

Many of the authors and interviewees emphasized the terms collaboration, automation, orchestration, integration and so on while expounding their definition of DataOps. For instance, "For DataOps to be effective, it must manage collaboration and innovation" [7], "DataOps is an analytic development method that emphasizes communication, collaboration, integration, automation, measurement and cooperation between data scientists, analysts, data/ETL (extract, transform, load) engineers, information technology (IT), and quality assurance/governance" [4], "Collaboration is the main part of both DevOps and DataOps" [3]. When describing their understanding of DataOps, most of the experts and authors elaborate their definitions with DataOps components and set of goals. Data pipelines to better explain the flow of data through operations [7], [29], [4], [10], the process of orchestration and automation. After analyzing the definitions, it was found that different definitions of DataOps seem to take different perspectives. While some focus on the activities of DataOps some focus on the goals of DataOps. Some focus on the technologies involved while some focus on the organizing structure of teams and so on. Tables 2 and 3 below categorizes the definitions we analyzed from interview studies and literature respectively.

From tables 2 and 3, it can be observed that some elements are common in all the definitions. Another important insight is that many terms associated with DataOps are also common to DevOps, Agile development and Big Data Analytics. We are also trying to identify the components/factors which make DataOps different from the others. We analyze the principles, goals, tooling, and people involved in all these different approaches/practices and formulate a definition for DataOps.

Definition for DataOps:

> *"DataOps can be defined as an approach that accelerates the delivery of high-quality results by automation and orchestration of data life cycle stages. DataOps adopts the best practices, processes, tools and technologies from Agile software engineering and DevOps for governing analytics development, optimizing code verification, building and delivering new analytics thereby promoting the culture of collaboration and continuous improvement."*

Even though DataOps has similarities with DevOps, agile methodology and big data analytics, it is still different from these existing approaches. It is a process-oriented approach to data that spans from the origin of ideas to the creation of graphs and charts which creates value.

DevOps merged Development and Operations teams to promote continuous integration and continuous delivery. Similarly, DataOps merges two data pipelines namely value pipeline and innovation pipeline. Value pipeline is a series of activities that produces value or insights and innovation pipeline is the process through which new analytic ideas are introduced in the value pipeline. In DevOps, the focus is on code and in data analytics, the focus should be both on code and data at every step. Moreover, DataOps has to deal with people along with tools due to which it requires a combination of collaboration and innovation.

Table 2: Analysis of DataOps definitions from literature

Perspective	Definition
activities	DataOps is not the product. It is an enabler of success [17]
activities	enables data analytics teams to thrive in the on-demand economy [6]
activities	help data teams evolve from a an environment with data silos, backlogs, and endless quality control issues to an agile, automated, and accelerated data supply chain that continuously improves and delivers value to the business [10].
way of working	DataOps is more than DevOps for data analytics because the deployment of a data pipeline is not a use case by itself [9]
way of working	focus on improving the communication, integration and automation of data flows between data managers and consumers across an organization [2]
way of working	works on data Management practices and processes which improves the accuracy of analytics, speed and automation [15]
way of working	DataOps uses technology to automate data delivery with the appropriate levels of security, quality and metadata to improve the use and value of data in a dynamic environment [14]
goal	The goal of DataOps is to create predictable delivery and change management of data, data models and related artifacts [13]
goal	bring rigor, reuse, and automation to the development of data pipelines and applications [10]
goal	By adopting DataOps, organizations can deliver data products in a consistent, reliable, fast, scalable, and repeatable process just like a factory [17]

DataOps and Big Data analytics are the two terms used interchangeably. However, DataOps is not only for Big data, instead it can be applied to any size of data to improve quality, speed and reliability of data insights.

In agile methodology, innovation happens in regular intervals. DataOps adopts this from Agile and as a result data team publishes new or updated analytics which is pushed into the value pipeline. Instead of copying best features of different approaches, DataOps borrows the best practices, technologies and tools and hand tailor it so that it fits to the unique context of data analytics.

Our definition of DataOps mostly aligns with the definition formulated by Ereth in [20]. In our definition, we call DataOps as a data strategy, because it sets the basis for transformation. Data strategy is something that is required by all organizations who make use of data for analytical purposes.

4.2 Use cases at Ericsson

Representatives of teams using raw data for developing data analytics, data engineers and data scientists were interviewed. The sections below describes the activities that they perform at Ericsson.

Table 3: Analysis of DataOps definitions from interviews

Perspective	Definition
activity	a process which fills the gap between data and operations team
goal	approach to eliminate data silos by connecting different data pipelines
goal	method to showcase the interdependence of end to end data analytic process
goal	reduce the risk of poor data quality and exposure of sensitive data that may cause problems for the organization
activities	enables a continuous and dissipated flow of access to and insights from data
activities	automate the build of pipeline environments and give data pipeline developers self-serve ability to create, test, and deploy changes
way of working	intersection of advanced data governance and analytics delivery practices that constitutes the data life cycle.
way of working	is a way of avoiding common mistakes organizations make in data science and analytics
way of working	connects data creators with data consumers to increase collaboration and digital innovation.
way of working	an efficient way of managing the activities in the entire data life cycle
way of working	brings together the suppliers and consumers of data thereby escaping from a static data lifecycle

Case A: Automated data collection for data analytics - Ericsson delivers software every second weekend to base stations located different parts of the world and data is collected from all of the base stations that are used on the continuous integration flow. Thus, several base stations run test cases 24/7. When the test case fails, it immediately generates some data, specifically test case related metadata and the log from the base station. This data is sent to the cluster where it is ingested, unzipped, packaged, and so on. The data thus collected is further utilized for performing software data analytics.

Case B: Building data pipelines - Data pipelines are built for easier production of insights from raw data which is collected from the devices. With the usage of data pipelines, the entire data process starting from origin of ideas to literal creation of charts can be done with a minimum human involvement. The execution of different stages of the pipeline can be controlled by the scheduler which triggers the execution of one job immediately after finishing the current one. To manage the evolving customer requirements, underlying code for the data pipelines are kept scalable. The data pipelines can be either same or different for different customers depending on the similarities in their requirements.

Case C: Toolkit for Network Analytics - Network analytics utilizes different types of network data collected from the devices out in the field to identify interesting and useful trends and patterns. This internal toolkit can monitor, analyse and troubleshoot networks automatically whenever an equipment fault is found. After the development of this toolkit, Engineers are able concentrate on high value tasks, consultant requirement got reduced and it shows a conservative saving of man-hours. This toolkit produces professional reports for the customers and enables new opportunities by providing real-time and historical data. Whenever the schema of the input data changes, then the pipelines will not take it and this scenario requires human intervention.

Case D: Building CI pipelines for Data Scientist teams - The targeted customers for this case are data scientist teams who make use of hardware analytics for predicting the quality of the hardware delivered to the customers. When the customers sent their product to the screening centers or the repair centers of Ericsson, the data gets recorded. The data scientist teams are collecting data from these centres to develop hardware analytics. The results or insights produced from the data can be used for machine learning algorithms for different activities. For instance, to predict if the customer is going to return the product or when the customer is going to return the product. This use case deals with building continuous integration pipeline for this data scientist team so that they get the feedback data continuously from the customers which can reduce the time for doing analytics as the data scientist team can get the data continuously. Apart from that CI pipeline will have basic unit tests and data linting tests.

Case E: Tracking the software version - To shorten the cycle time towards the customers there should be feedback loop from the customers. However, that's been very difficult, as the customers are in other countries and different companies. In order get data back from the customers software version running at the customer site needs to be tracked. Every third week software is delivered to the customers and then follow up is done to check which software they're unning, and also collect some performance data to ensure that the networks are performing adequately. Apart from the data at customer side, there is also data from internal CI environments which requires follow up. If an issue occurs in a lower level testing context it can be seen in high level testing or vice versa. So, the fingerprint of a certain issue can be seen across all the test levels. It's quite important to relate these issues. Otherwise there can have a bug which appears with ten different symptoms in ten different test environments, and it's difficult to debug. And the third is the customer data. The customers typically have a very good knowledge of their networks, they're very skilled at analytics, also. But, sales departments might not have same skills. Thus, it is required to help those departments understand data by creating dashboards out of data.

Case F: Testing the software quality - Features of the deployed software and those that are planned to be released in the future releases needs to undergo software quality tests with the help of KPIs. KPIs formulated will check if the system introduced on the software are reflecting what is expected as per design. This applies for the upgrades in the features as well. Data collected from the customers like counters are used to formulate KPIs. KPIs are used for monitoring if the feature behavior is as expected or as designed. If the KPIs are following the usual trend, then the performance is as expected. There are two different tools which help in KPI monitoring.

Case G: KPI analysis Software - KPI analysis software helps to turn the KPI analysis into informed business decisions. KPI analysis is performed on the nodes in the continuous deployment zone

before and after product updation. There is a mechanism to collect data automatically from the nodes. After getting the data, KPIs are defined manually and given to the software which then learns the trends of the counters or the KPIs and calculates it in the baseline. Once the updation happens, it again collects data from the devices and does the same again, but this time it compares the newly learnt trend with the baseline and the result of this will be charts or graphs representing performance improvements or performance degradation. These insights are delivered to the customers in an agile fashion. i.e every third week. After deploying the software, the team continuously monitor the data from the nodes.

Case H: Building data pipelines for CI and CD data - Building data pipelines for Continuous Integration and Continuous deployment enables access to data for all the team working with analytics. The main objective of this use case is to provide availability of high quality data to all the teams who are using data. A data pipeline with 4 steps such as data ingestion, data downloading, data archiving, data processing and data serving is built and it is manually monitored continuously to check for the data quality and data availability. Whenever there is a variation from the usual pattern, immediately the person responsible for the pipeline robustness is informed and that person finds the reason for the error and fixes it. So, for this particular use case, monitoring part involves human intervention as most of the tasks are done manually.

4.3 Evolution of DataOps

Based on the cross case analysis and literature, we identified a five stage evolution which happened before the introduction of DataOps. Because, the cases described above were not built in a single stretch to implement DataOps. Rather, they were built over time without even knowing that these would become beneficial in the future. From this inference, we thought of developing an evolution model with different stages that the company has gone through. Each of the cases mentioned above is developed at different stages. When climbing the stairway of evolution model, these cases/components are either taken as such or necessary modifications are performed to take it further to the next stage. There might have other components as well in each of the stages. However, we are not considering all of those components. Instead, we consider those cases which are developed at some stage and taken over to the successive stages.

The stairway shown in fig 2 depicts the different maturity levels or evolution stages of data collection wherein data was initially collected in an ad-hoc fashion and progressing to completely autonomous unit that collects data, does data analytics, monitors itself for anomalies thereby reducing the time for delivering insights. The cases studied at Ericsson are mapped as components used at each of the evolution phases.

This evolution of stages in the taxonomy occurs on a component basis. Essentially, data life-cycle activities (*data collection, data preparation, data analysis, and delivering insights*) starting from origin of ideas to literal creation of values in the form of charts and graphs are performed at all maturity stages. The four phases of the "DataOps Evolution Model", namely "Ad-hoc data collection", "Data Pipelines and Data technologies", "Agile Data Science", "Continuous testing and Monitoring" and "DataOps", are described in detail in the remainder of this section.

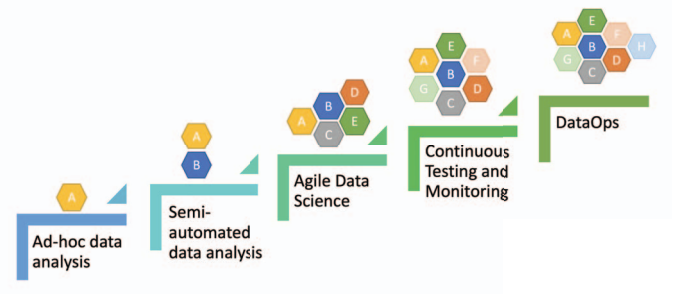

Figure 2: Evolution of DataOps

Phase 1: Ad-hoc Data Analysis

In ad-hoc data analysis, the reports or insights are created on-demand due to which the reports were highly customized. Usually, an ad-hoc analysis was performed to answer a very specific business question. The ad-hoc analysis was used in the early days of Business analytics. This was highly dependent on the templates provided by the IT department. An ad-hoc analysis lets the user decide which data sources to fetch from and the way of data presentation. Ad-hoc insights can range from simple one-off data table all the way to intricately detailed sales reports using dashboards, interactive maps, and other advanced visualization features. Another important feature of ad-hoc data analysis was its ability to deal with different data sources in a flexible and scalable way. The ad-hoc analysis is helpful when there is a requirement of delivering immediate results. However, the reports generated out of this analysis are not used after the intended purpose. To make good business decisions, it is always better to have proper data engineering, data collection, and extensive data analysis. One of the practitioners commented that

> *"Because the data collection is basic plumbing, you know? You're moving one bit from one place to the other, you have to set up how the flow of data that goes from one end to the other. But, fully automated data analysis is something that we initially struggled with."*

Requirements for this phase:

To do ad-hoc data analytics, there should be some technology with which real-time data can be collected from multiple data sources.

Challenges:

All the data collected from different sources would not be in a single access point. Data silos at this phase prevent the customer from getting the full picture. Business decisions rely on a small amount of data which is not often sufficient to make decisions. Improper handling of an organization's data can lead to conflicts in the business values developed. When the underlying data varies throughout the organization, it can lead to conflicting results and delayed decisions.

Phase 2: Semi-Automated data analysis

Data pipelines for collecting and processing data are a much more efficient and automated way to implement data analytics. One of the interviewees said that

> *"It is more complex to build a robust data pipeline that is robust, reusable, scalable, secure and traceable in a*

real-world scenario. Data pipeline which we have built right now is reusable, secure and traceable, but we are not quite sure about its scalability and robustness."

With the advent of data pipelines, data technologies and data processes became a necessity as they control and co-ordinate the different phases of data pipelines.

Data Pipelines: The huge volume of raw data is generated through various sources both internal and external to Ericsson. Data generated by different teams contribute to Internal sources and data generated by the devices at the base stations contributes to the external sources. Ericsson follows a similar data pipeline as explained in [24]. Data is collected in the form of raw dumps. Considering the complexity, heterogeneity, and volume of big data, Ericsson executes its applications in various stages as described below.

From figure 3, it can be understood that the data pipeline shown is a value pipeline as it creates visualization or insights from the collected raw data. According to the use cases, the data pipeline changes its face while the basic steps or structure being the same.

Data Technologies: Data pipelines require technologies to ingest, clean, analyze and visualize data. The technologies used to manage the pipeline can be categorized into four namely Data Engineering, Data Preparation, Data Storage and Data Visualization.

A. Data Engineering: The Data Engineering step performs two different operations at a high level, which include data collection and data ingestion. The process starts with continuous data streams collected from multiple sources including internal as well as external sources and ingested into the data pipeline. The data ingestion is important because data ingestion method itself is different for different data sources. For instance, the data ingestion method used for ingesting CI (Continuous Integration) data collected from internal sources is different from the ingestion method used for ingesting CD (Continuous Deployment) data collected from the external sources. Data ingestion is capable to collect, import and process data from different data sources.

B. Data Preparation: Despite the collection of highly relevant data, analytics should take into account data heterogeneity to maintain efficiency in real-time applications. Data preparation involves the preparation of metadata links to the path where the actual data is stored and aggregating all the links for different types of data. Identification of encoded/encrypted data takes place here and once these kinds of encoded messages are identified, message to decode it is sent to the third-party servers. Encoded data are decoded by third-party servers and the respective metadata links are sent back to the aggregation module.

C. Data Storage: Metadata links prepared by the data preparation module is then stored in the Hadoop database. Teams can search for the metadata links in the database and can download the raw data dump files through the downloader. There are two different databases - CFIDB and Hadoop Database where the storage of data happens. CFIDB stores the CI data initially when it is collected from the internal teams at Ericsson. Hadoop database is the main database where the storage of aggregated data-logs takes place.

D. Data Visualization: After preparation and storage, the process of data analytics is executed. According to the requirement, different teams at Ericsson access the downloader to download the raw data dumps from the Hadoop database. After downloading the raw data dumps from the database, steps like data cleaning, data filtering, data processing, data transformation, etc are performed according to the requirements of the stakeholders. Most of the stakeholders require reports on the data showing the performance variation after the installation of a particular device.

Requirements for this phase: Well designed data pipelines are needed to efficiently evaluate, test, ingest, transform, validate, and publish data at scale. Data technologies for data collection, data engineering, data processing, data analysis, and data visualization. Also, data processes to control and coordinate data technologies as well as data pipelines are required

Challenges Lack of data pipeline robustness. Some of the activities are not automated. For instance, monitoring is done manually and whenever some issues are found, dependencies need to be contacted manually to fix the issue. The tickets raised while encountering problems in the data pipeline takes too long to get fixed. Once the insights are delivered, the process basically stops. Feedback from the customers is not collected for further improvement.

Phase 3: Agile Data Science

Development and deployment are well defined by agile and DevOps methodologies. With these, teams are able to develop fully tested, functional code in a very short duration. Teams store their work in a common central repository in order to synchronize. There are a number of tools to aid the development and deployment phases. Customer requirements change with time and in order to cope up with the evolving requirements, it is required to follow the agile methodology in which insights are delivered in short sprints. Ericsson has a three weeks sprint means insights are delivered to the customers every third week, gets feedback from them and rework if required. A major challenge identified is that most of the time, stakeholders are not quite sure about their requirements. One of the interviewees quoted that

"A lot of the time you might create a dashboard or service which the stakeholders think they want, but at the end of the day, we see that it's almost never used. It usually takes some back and forth before we're able to find that killer app for the stakeholder."

Requirements for this phase: Continuous delivery of business values to the customers. Evolving requirements from customers should be addressed. Customers should be delivered their demands frequently. Data team and customers should interact and the customers should communicate their requirements directly to the data team. Team should adjust themselves to increase the efficiency after regular intervals

Challenges Without continuous automated testing, a lot of man-hours are required to guard the flow in the data pipelines. To deliver insights quicker, good quality data should be made available. For instance, if the data source is not sending data, it should be detected as early as possible so that actions can be taken immediately. For this, data flowing through the pipelines should be monitored continuously.

Phase 4: Continuous testing and monitoring :

Continuous testing and monitoring of the data pipeline is an essential element while dealing with real-time data. Because it can help to detect the problems immediately before it is carried over to

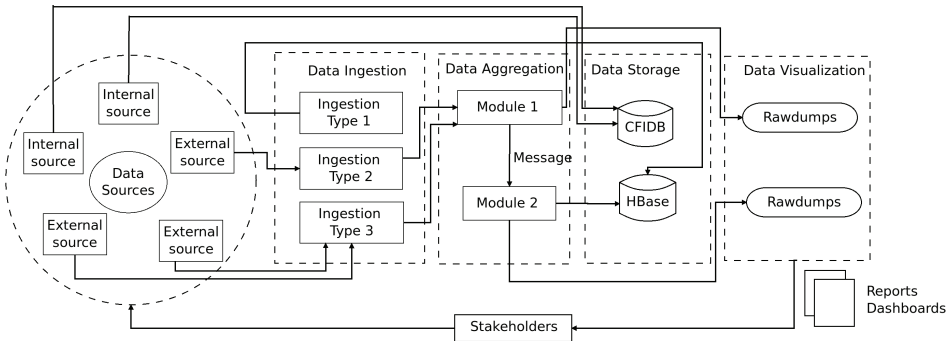

Figure 3: Big Data Analytics Pipeline at Ericsson

the successive stages of the pipeline. Without a monitoring mechanism, when the data received at the end of the pipeline is not as expected, it will be hard to identify the reason for the unexpected output. Also to meet the quality constraint, it is very important to have automated unit tests as well as higher-level testing. Unavailability of higher-level testing was quoted as a challenge by one of the interviewers and it goes like this

> *"We are deploying automatically, I think, in most cases, but we don't have the quality checks and balances that we need to have. So, you can push something which doesn't have adequate quality and the pipeline still accepts it."*

With automated alerts, the concerned team can be notified when something goes wrong with any of the stages in the pipeline or if the pipeline is broken and the team can take proper measures to ameliorate the effect of the breakage.

Requirements for this phase: Test cases for testing the quality of data flowing through the pipeline. Automated mechanism to perform continuous monitoring and automatic alerting mechanism to send alarm to the responsible team when encountered with pipeline issues. Mitigation strategies should be developed in order to handle pipeline breakage

Challenges When there are pipelines, there should be some way to manage and orchestrate operational characteristics of the pipeline. Mechanism to push new data analytic ideas into the existing value pipeline

Phase 5: DataOps

DataOps shortens the end-to-end cycle time of data analytics, from the ideation phase to the insight development. As data lifecycle has dependency on people in addition to tools, it incorporates Agile Development practices into data analytics according to the organization's requirement thereby bringing the data consumers and data suppliers work together more efficiently and effectively. DataOps also adopts DevOps principles to effectively manage the artifacts like data, metadata and code. One major difference between DevOps for data analytics and DevOps for software development is that former has to manage both data and code whereas latter concerns only about the evolving code. With DevOps, it brings the two foundational technologies - continuous integration and continuous delivery which are two essential factors contributing to the goals of DataOps [7], [10], [11].

In addition to the DevOps lifecycle, data lifecycle has got an intersection between two pipelines namely value pipelines and innovation pipelines. Value pipelines are used for the creation of insights and innovation pipeline is for injecting the new analytic ideas into the value pipelines. In DataOps beyond the automated deployment of infrastructure, software, and application code, there is the requirement of orchestration. The data pipeline starting from from acquisition of raw data to development of data product typically but not always follows a directed acyclic graph. DAG data structure has nodes and edges connecting the nodes. Nodes are the tasks where data is stored and edges denotes the flow of data from one node to another. The edges are directed because data cannot flow in the opposite direction. The output of one task becomes the input for another. The DAG is always acyclic because moving from node to node will never create an edge to a previous node. As the execution of steps occurs in a specific sequential order respecting the dependencies between different components, DAG usually requires orchestration. However, with the rise in real-time streaming architectures choreographed DAGs are becoming more popular. Because of the above mentioned reasons, automation, Orchestration, collaboration are the most important elements of DataOps.

Ericsson wants to apply DataOps to accelerate the data analytics workflow. At this point, the organization is heading towards the last step of the evolution of the stairway, which is DataOps. At least there is an initiative to organize all the people who work on data as a team so that data silos can be reduced. Moreover, with this initiative, all the teams associated with data can get to know what the other team is doing which makes the whole process of data analytics better. DataOps requires this sort of reorganization of the teams along with the value pipeline and innovation pipeline. However, there are concerns regarding all the data teams downloading data from the same place. Because the existing pipeline might not be able to serve a larger number of data requests.

There are several value pipelines created according to the requirements from the customers. However, all these pipelines share a common skeleton. Although, there is innovation pipeline, it is not much established and it is hard to explain how the new analytic ideas are pushed into the value pipeline.

Requirements for this phase: Data pipelines for creating insights and innovation pipelines for pushing new analytics into data pipelines. Continuous integration and continuous delivery practices for data analytics. DevOps for Data analytics. Mechanism to

monitor and control the entire data life cycle process. Orchestration and advanced automation and agile practices for data analytics

Challenges Organizational restructuring is required. Unavailability of skilled team proficient in both Data analytics and DevOps is another challenge. Lack of interest of data scientists in learning new tools and technologies and data silos are the other major impediments.

5 THREATS TO VALIDITY

There are three categories of potential threats to the validity of our work. This include construct validity, reliability and external validity that needs to be taken into consideration. To ensure construct validity, a few cases were excluded from the results as some of the interviewers did not had proper understanding of DataOps. As a result of the screening process, our study have some limitation with number of interviews. However, this limitation can be counted as an opportunity for further inquiry in future works. For reducing the researcher bias, the interviews were conducted by two researchers. To minimize internal validity threats, one of the co-authors, who has in-depth knowledge about the data processes in the company, was asked to validate the findings. Also, the findings were validated with other employees at the company. External threats we foresee are how the findings can apply to other organizations. Moreover, our reliance on grey literature as data sources for analysis also serves as a limitation. Further validation can be done by involving more organizations, which we see as future work.

6 CONCLUSION

DataOps is becoming increasingly popular in the industry due its ability to accelerate the production of high quality data insights. This paper proposes an evolution model describing a strairway with five steps showing how DataOps was evolved. With our research contribution, which is based on an extensive case study at Ericsson, we aim to provide guidance on this topic and enable other companies to establish or scale their DataOps practices. Our main contribution is the "DataOps Evolution Model". In the model, we summarize the five phases of evolution and maps cases to the phases in which they are used. Researchers and practitioners can use this model to position other case companies and guide them to the next phase by suggesting the necessary features. As future research, we plan to validate our model with other companies.

7 ACKNOWLEDGMENTS

This work is in part supported by Vinnova, by the Wallenberg Artificial Intelligence, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation and by the Software Center. The authors would also like to express their gratitude for all the support provided by Ericsson.

REFERENCES

[1] [n.d.]. 3 reasons why DataOps is essential for big data success | IBM Big Data & Analytics Hub. https://www.ibmbigdatahub.com/blog/3-reasons-why-dataops-essential-big-data-success. (Accessed on 01/24/2020).
[2] [n.d.]. Data Ops. https://www.gartner.com/en/information-technology/glossary/data-ops. (Accessed on 01/14/2020).
[3] [n.d.]. DataOps - Devops for Big Data and Analytics | XenonStack. https://www.xenonstack.com/insights/what-is-dataops/. (Accessed on 01/14/2020).
[4] [n.d.]. DataOps and the DataOps Manifesto - ODSC - Open Data Science - Medium. https://medium.com/@ODSC/dataops-and-the-dataops-manifesto-fc6169c02398. (Accessed on 01/14/2020).
[5] [n.d.]. DataOps: Changing the world one organization at a time | ZDNet. https://www.zdnet.com/article/dataops-changing-the-world-one-organization-at-a-time/. (Accessed on 01/14/2020).
[6] [n.d.]. DataOps in Seven Steps - data-ops - Medium. https://medium.com/data-ops/dataops-in-7-steps-f72ff2b37812. (Accessed on 01/25/2020).
[7] [n.d.]. DataOps is NOT Just DevOps for Data - data-ops - Medium. https://medium.com/data-ops/dataops-is-not-just-devops-for-data-6e03083157b7. (Accessed on 12/20/2019).
[8] [n.d.]. The DataOps Manifesto. https://www.dataopsmanifesto.org/. (Accessed on 01/14/2020).
[9] [n.d.]. DataOps: More Than DevOps for Data Pipelines. https://www.eckerson.com/articles/dataops-more-than-devops-for-data-pipelines. (Accessed on 01/25/2020).
[10] [n.d.]. Diving into DataOps: The Underbelly of Modern Data Pipelines. https://www.eckerson.com/articles/diving-into-dataops-the-underbelly-of-modern-data-pipelines. (Accessed on 01/14/2020).
[11] [n.d.]. The Emergence of DataOps Empowers the Future of Data Management | Analytics Insight. https://www.analyticsinsight.net/emergence-dataops-empowers-future-data-management/. (Accessed on 01/24/2020).
[12] [n.d.]. From DevOps to DataOps - DataOps Tools Transformation | Tamr. https://www.tamr.com/blog/from-devops-to-dataops-by-andy-palmer/. (Accessed on 01/14/2020).
[13] [n.d.]. Get Ready for DataOps - DATAVERSITY. https://www.dataversity.net/get-ready-for-dataops/. (Accessed on 01/25/2020).
[14] [n.d.]. What is DataOps? - DataOps zone. https://dataopszone.com/what-is-dataops/. (Accessed on 01/25/2020).
[15] [n.d.]. What is DataOps? Everything You Need to Know | Oracle Data Science. https://blogs.oracle.com/datascience/what-is-dataops-everything-you-need-to-know. (Accessed on 01/25/2020).
[16] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 291–300.
[17] Harvinder Atwal. 2020. The DataOps Factory. In *Practical DataOps*. Springer, 249–266.
[18] Pamela Baxter and Susan Jack. 2008. Qualitative case study methodology: Study design and implementation for novice researchers. *The qualitative report* 13, 4 (2008), 544–559.
[19] Jan Bosch. 2017. *Speed, data, and ecosystems: Excelling in a software-driven world*. CRC press.
[20] Julian Ereth. 2018. DataOps-Towards a Definition.. In *LWDA*. 104–112.
[21] Vahid Garousi, Michael Felderer, and Mika V Mäntylä. 2016. The need for multivocal literature reviews in software engineering: complementing systematic literature reviews with grey literature. In *Proceedings of the 20th international conference on evaluation and assessment in software engineering*. ACM, 26.
[22] Vahid Garousi, Michael Felderer, and Mika V Mäntylä. 2019. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and Software Technology* 106 (2019), 101–121.
[23] Sylvia Ilieva, Penko Ivanov, and Eliza Stefanova. 2004. Analyses of an agile methodology implementation. In *Proceedings. 30th Euromicro Conference, 2004.* IEEE, 326–333.
[24] HV Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M Patel, Raghu Ramakrishnan, and Cyrus Shahabi. 2014. Big data and its technical challenges. *Commun. ACM* 57, 7 (2014), 86–94.
[25] Lucy Ellen Lwakatare, Pasi Kuvaja, and Markku Oivo. 2016. An exploratory study of devops extending the dimensions of devops with practices. *ICSEA 2016* 104 (2016).
[26] Moira Maguire and Brid Delahunt. 2017. Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars. *AISHE-J: The All Ireland Journal of Teaching and Learning in Higher Education* 9, 3 (2017).
[27] Rodney T Ogawa and Betty Malen. 1991. Towards rigor in reviews of multivocal literatures: Applying the exploratory case study method. *Review of educational research* 61, 3 (1991), 265–286.
[28] Per Runeson and Martin Höst. 2009. Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering* 14, 2 (2009), 131.
[29] Prabin Ranjan Sahoo and Anshu Premchand. 2019. DataOps in Manufacturing and Utilities Industries. (2019).
[30] Mojtaba Shahin, Muhammad Ali Babar, and Liming Zhu. 2017. Continuous integration, delivery and deployment: a systematic review on approaches, tools, challenges and practices. *IEEE Access* 5 (2017), 3909–3943.